Content validity of measures of theoretical constructs in health psychology: discriminant content validity is needed

Introduction

Many of the theoretical constructs and outcomes of interest to health psychology cannot be objectively assessed. For example, phenomena such as beliefs, pain, health, quality of life, stress, intention, illness representations are all of interest but none are available for direct measurement. Rather, the measurement of such theoretical constructs is an inferential process requiring the development of instruments that assess the target construct indirectly, typically using questionnaire based measures. Establishing and reporting the psychometric properties of such measures is challenging but fundamental to their utility in testing theory, designing and evaluating interventions and making clinical and policy decisions.

The psychometric assessment of measures of these health related constructs, including health outcomes and predictors of health outcomes, has advanced in terms of reliability (the degree to which scores on a measure are consistent) and some aspects of construct validity . By contrast, current conventions of reporting typically omit content validity. Examination of the history of the concept of validity sheds some light on why content validity has come to be neglected. Up until the 1980's the APA Standards for Educational and Psychological testing adopted a tripartite approach to (test) validity, namely, criterion, content and construct related validity (Sireci & Sukin, 2013). The current APA definition of each form, adapted to be relevant to health psychology, is given in Table 1. However, this tripartite approach was replaced by a unitary conceptualisation of validity in which construct validity subsumed all other aspects, categories or types of validity. Although the proponents of the unitary conceptualization of validity recognised the importance of representative and relevant content they nonetheless argued that representative and relevant content is not a form of validity. This argument has, over time, likely led to a neglect of content validity. However, the unitary conceptualization, has never been universally accepted. Indeed, some predicted that the refusal to accept content validity as a form of validity would eventually be detrimental to validation practices (Sireci, 1998; Yalow & Popham, 1983); we share that view.

Table 1 about here

In addition, the lack of agreed transparent methods of assessing and reporting content validity also contribute to its neglect. Thus it is possible to report *'how well'* a measure is performing without being able

to specify *'what'* it is measuring. Here we argue that the explicit evaluation of content validity would enable researchers and clinicians to select existing measures that truly assess what they aim to measure without ambiguity, overlap or contamination from other related constructs and, where no such measures exist, enable them to develop new content valid measures. By adopting a convention of reporting content validity the pitfalls of poor content validity might be avoided and methods of establishing and assessing content validity might be improved.

Defining content validity and discriminant content validity

Content validity refers to "*the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular purpose*" (Haynes, Richard, & Kubany, 1995). Content validity is considered the most important aspect of a measure of a theoretical construct (Terwee et al., 2018). Content validity is fundamental as it specifies what is being measured. Establishing the content validity of a measure requires that both components, relevance and representativeness, be assessed. First, does the measure accurately reflect the focal construct, i.e. the theoretical construct it aims to measure; do the items that form the measure have relevance to the focal construct? Second, does the measure reflect the whole breadth of the focal construct; is the measure representative of the whole construct? Thus content validity is determined by the relationship between the definition of a construct and the items designed to measure it.

Content validity has the potential to influence the interpretation of all other psychometric properties of a measure, including construct validity and reliability. Most obviously, if a measure is found to be highly reliable but has poor content validity, the interpretation of a score will be entirely erroneous because the measure has no meaning in relation to the target construct.

Construct validity requires that a measure functions as the proposed construct does, but it is possible to achieve construct validity without content validity. For example, in testing the relationship between intention and activity, a measure might predict activity in a manner consistent with theory without containing any intention content, e.g., if it sampled other cognitions related to intention such as attitudes toward activity. And, whilst it is important to establish the internal structure of a measure (Crutzen & Peters, 2017), finding evidence that a measure has the hypothesised factor structure does not in itself demonstrate

content validity. Although authors may choose to 'name' the factors to match the intended constructs, the factor may simply contain items which assess the determinants or consequences of the intended construct.

Even when a measure has content validity, it may not have *discriminant content validity*, i.e. content that is distinguishable from the content relevant to other constructs. This can be a significant problem where there are closely related or overlapping constructs. For example, self-efficacy and perceived behavioural control are similar constructs from different theories, but examination of the content validity of measures of these constructs found items which related to neither definition and even one item purporting to measure perceived behavioural control but instead measuring self-efficacy (Johnston et al., 2014). Similarly, measures of pain-catastrophising were considered to have good construct validity due to their performance in predicting activity limitations, but a recent analysis of the content of six standard measures suggest that they lack discriminant content validity. The measures did not adequately reflect the definition of pain catastrophizing and were a better fit to other pain constructs including 'pain-related worrying' or 'pain-related distress' (Crombez, De Paepe, Veirman, Eccleston, & Van Ryckeghem, in submission).

Importance of (discriminant) content validity for theory, intervention design and practice

Content validity is important, for the testing of theory, for the design of behaviour change interventions and for the measurement of health outcomes of importance to patients and clinicians. A lack of content validity weakens theory testing and the interpretation of data.

Valid tests of theory depend on having measures with discriminant content validity, otherwise apparent relationships may simply be due to measurement confound. Contamination of a measure by content relevant to a related construct is particularly problematic when the measures are used to examine relationships between the two constructs. For example, many studies of people with musculoskeletal conditions examine the theoretical relationship between impairments, such as pain, and activity limitations, such as limitations in the ability to walk or get dressed. However, existing outcome measures typically have a mixture of content embracing both impairment and activity limitations (Pollard, Johnston, & Dieppe, 2006). Thus any relationships found may simply be due to the contaminating content, i.e., lack of discriminant content validity in the measures. When pure, uncontaminated measures are used, the relationships are considerably diminished or non-existent (Pollard & Johnston, 2001).

Knowing what one is measuring may be crucial in designing an intervention. If measures lack content validity but are otherwise psychometrically sound, then they may lead to mis-identification of determinants of the target mood, behaviour or symptom. Recent work has identified behaviour change techniques that are appropriate for targeting key theoretical constructs that act as mechanisms of action to determine change in the target behaviour (Carey et al., 2018; Connell et al., 2018; Johnston et al., 2018). Clearly success in making use of such evidence depends on valid labelling of the theoretical construct and this process may be derailed by misleading classification of measures. For example, some self-efficacy items have been found to tap 'motivation' as well as self-efficacy (Burrell, Allan, Williams, & Johnston, 2018) and if used as a basis for designing an intervention might result in the selection of less appropriate techniques than might be achieved if determinants of the behaviour were identified using measures with content validity.

In practical applications, test scores are used to make decisions, including planning and policy decisions that affect availability of resources, as well as decisions about clinical interventions. If hospital managers attempting to reduce work stress in nurses used current measures to assess factors influencing work stress, they might be misled about the key determinants, such as workers' control, since the main measure assesses both control over decisions and use of skill and individual items have poor content validity (Bell, Johnston, Allan, Pollard, & Johnston, 2017). A clinician might reach a different treatment decision depending on the content validity of the measure used for patient assessment. Measures used to assess limitations in patients with arthritis range from those which contain items mainly assessing impairment to those which are mainly participation restrictions (Pollard et al., 2006); it is therefore unsurprising if they result in differing assessments of degrees of severity and if they differ in sensitivity to treatment effects of pharmacological and exercise interventions (Ayis et al., 2010).

Developments of digital and mobile technologies make the content validity problem more, rather than less, urgent. With the exception of direct observational and sensing measures, the commonly used EMA (ecological momentary assessment) methods rely on very frequent self-reporting of behaviours or of intrapsychic phenomena such as beliefs, emotions, symptoms and their precursors or consequences. In order to reduce participant burden, each construct is typically measured by a single or a few items and it is therefore particularly important that each of these items has content validity for its target construct.

The use of an agreed, feasible, evidence-based methodology for establishing essential aspects of content validity of construct measures, would considerably advance our ability to develop and/or select valid measures suitable for theory testing and intervention development, and the assessment of outcomes important to patients and clinicians.

Methods of assessing CV and DCV

Two methods are currently available for assessing content validity, namely, the Content Validity Index (CVI) (Lynn, 1986) and Discriminant Content Validity (DCV) (Johnston et al., 2014). While CVI has been used to assess content validity of health outcome measures, DCV has been used to assess the content validity of theoretical process variables (Bell et al., 2017; Burrell et al., 2018; Gardner, Abraham, Lally, & de Bruijn, 2012; Johnston et al., 2014) and theoretical domains (Cane, O'Connor, & Michie, 2012; Huijg, Gebhardt, Crone, Dusseldorp, & Presseau, 2014), as well health outcome measures (Dixon, Johnston, Mcqueen, & Court-Brown, 2008; Dixon, Pollard, & Johnston, 2007; Pollard et al., 2006; Schmitt et al., 2013).

Both are judgement tasks, whereby judges rate the extent to which measurement items are related to the target construct. Therefore, both CVI and DCV can be used in the development of a valid measure, prior to evaluation of reliability and construct validity by testing with participants. However, there are differences between the two methods. CVI examines the relevance of items in relation to a single target construct and is scored such that each item is judged either relevant or not. The proportion of judges in agreement about relevance is then used to determine the CVI for each item (I-CVI) and the CVI for the whole scale/measure (S-CVI) (Polit & Beck, 2006). However, it does not quantify the extent to which the measure is distinct and uncontaminated by other constructs.

Figure 1 about here

We developed the DCV method to provide that additional important information. Like the CVI, DCV establishes the content validity of items in relation to the target construct, but also establishes content validity in relation to other constructs, either from the same theory or related constructs from other theories. In addition, a DCV study also identifies items that *do not* measure the construct (see the illustration in Figure 1, where the item is judged, with high confidence, to be measure self-efficacy, but with more modest confidence to *not* measure perceived behavioural control). Thus, DCV establishes the discriminant content validity of items necessary for theory testing and for precise assessment of intervention effects. DCV data

also give a quantitative estimate of the content validity of items (Crombez et al., in submission; Johnston et al., 2014). Thus, DCV can be used to identify items that are pure measures of a single construct uncontaminated by other constructs and can examine whether an item measures one construct more strongly than others. It is then possible to choose items as required, for example, theory testing requires pure measures, whilst items measuring multiple constructs may be useful as single item measures, especially in clinical settings (Johnston et al., 2014).

Further challenges

Both CVI and DCV methods focus on 'relevance' and only one DCV study has attempted to develop a method for assessing representativeness (Bell et al., 2017), the other key component of content validity. Standard measures of work stress lacked items to assess important parts of work stress definitions, and also contained items that were not relevant to the definitions. Similar attempts to assess representativeness have been explored in organizational psychology (MacKenzie, Podsakoff, & Podsakoff, 2011).

Content validity studies have highlighted the importance of the definitions of the target constructs and others have noted this problem in establishing scale validity: "Given the importance of clearly defining the conceptual domain of the construct, it is surprising that so many researchers either neglect this step in the process or fail to properly implement it" (MacKenzie et al., 2011). If definitions are imprecise or simply lacking then it is virtually impossible to establish content validity. For example, there are no agreed definitions of work stress constructs and the content validity of items in a self-report measure of work stress differs when different definitions of the construct are used (Bell et al., 2017).

To date, content validity studies have tended to focus on theoretical constructs and health outcome measures with rather less emphasis on the content validity of other types of constructs, such as process variables and other outcomes of interest to health psychology, for example, self-reported measures of adherence, dietary or other health behaviours. Most of the work has been done on the content of the question and less on the content of the response format: for example, evaluating frequency versus intensity versus agreement formats. Future work might usefully begin to apply content validation methods more widely.

In addition, both CVI and DCV are judgment tasks, yet there is no agreement as to how judges should be selected. Typically, expert judges are used but should judges be experts familiar with the theoretical constructs or should they be drawn from the population of participants who will be the

respondents when the measure is used? The performance of expert judges on a DCV task were found to have clearer discriminations of items assessing illness perceptions than lay respondents similar to the intended respondents using the measures (Glidewell, 2008). Thus items that might have CV for the expert judges might lack validity for lay judges. This raises the issue of synthesis across content validity studies. There is surely merit in replication studies that examine the content validity of the same items or measures using different samples of the same type of judge or across different types of judges. Suitable methods of data aggregation could then be used to improve the reliability of content validity measures.

Finally, neither method overcomes the need for the robust methods to initially establish the pool of items relevant to, and representative of, the target construct. Qualitative methods including focus groups, interviews, and cognitive interviews with the target population have been used, together with expert opinion to improve the range and intelligibility of items (Prinsen et al., 2016).

Conclusions

Good measurement of key theoretical constructs is fundamental to much research and application in health psychology and we typically present careful assessments of aspects of reliability and validity. However we neglect the need to demonstrate the content validity of our measures. Without satisfactory content validity and discriminant content validity results may be meaningless and worse, may lead to erroneous conclusions in testing theory, choosing interventions and making clinical and policy decisions.

In part this deficit may be due to lack of clearly established methods, but CVI has been available for many years. It would therefore appear that we have simply established a convention of omission of content validity as part of the reporting of psychometric properties. There has been some recognition of the confusion of overlapping, ambiguous and confounding of labels for theoretical constructs but we have avoided contemplation of the hazards of failure to establish the discriminant content validity of our measures. We suggest that in future authors, reviewers and editors might seek better evidence of content validity (and especially discriminant content validity) of measures used in empirical studies and that research continues to address the challenges of improving content validity assessment.

References

Ayis, S., Arden, N., Doherty, M., Pollard, B., Johnston, M., & Dieppe, P. (2010). Applying the Impairment, Activity Limitation and Participation Restriction Constructs of ICF (WHO) Model to OA and low back pain trials: re-analysis *Journal of Rheumatology, 37*, 1923-1931.

Bell, C., Johnston, D., Allan, J., Pollard, B., & Johnston, M. (2017). What do Demand-Control and Effort-Reward work stress questionnaires really measure? A discriminant content validity study of relevance and representativeness of measures. *British Journal of Health Psychology, 22*, 295-329.

Burrell, A. M. G., Allan, J. L., Williams, D. M., & Johnston, M. (2018). What do self-efficacy items measure? Examining the discriminant content validity of self-efficacy items. *British Journal of Health Psychology*, 23: 597-611.

Cane, J., O'Connor, D., & Michie, S. (2012). Validation of the theoretical domains framework for use in behaviour change and implementation research. *Implementation Science, 7*, 37.

Carey, R. N., Connell, L. E., Johnston, M., Rothman, A. J., de Bruin, M., Kelly, M. P., & Michie, S. (2018). Behavior Change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. *Ann Behav Med*. kay078, https://doi.org/10.1093/abm/kay078.

Connell, L. E., Carey, R. N., de Bruin, M., Rothman, A. J., Johnston, M., Kelly, M. P., & Michie, S. (2018). Links between behavior change techniques and mechanisms of action: an expert consensus study. *Ann Behav Med*. doi:10.1093/abm/kay082

Crombez, G., De Paepe, A. L., Veirman, E., Eccleston, C., & Van Ryckeghem, D. (in submission). Let's talk about pain catastrophizing measures: an item content analysis. *Pain*.

Crutzen, R., & Peters, G. Y. (2017). Scale quality: alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychol Rev, 11*, 242-247.

Dixon, D., Johnston, M., Mcqueen, M., & Court-Brown, C. (2008). The Disabilities of the arm, shoulder and hand questionnaire (DASH) can measure the impairment, activity limitations and participation restriction constructs from the international classification of functioning, Disability and Health (ICF). *BMC Musculoskeletal Disorders, 9*.

Dixon, D., Pollard, B., & Johnston, M. (2007). What does the chronic pain grade questionnaire measure? *Pain, 130*(3), 249-253.

Gardner, B., Abraham, C., Lally, P., & de Bruijn, G. J. (2012). Towards parsimony in habit measurement: testing the convergent and predictive validity of an automaticity subscale of the Self-Report Habit Index. *Int J Behav Nutr Phys Act, 9*, 102.

Glidewell, E. (2008). *How do people with a diagnosis, caregivers and healthcare professionals represent dementia: An exploratory assessment using the Common Sense Self Regulation Model.* University of Aberdeen, PhD Thesis.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*, 238-247.

Huijg, J. M., Gebhardt, W. A., Crone, M. R., Dusseldorp, E., & Presseau, J. (2014). Discriminant content validity of a theoretical domains framework questionnaire for use in implementation research. *Implementation Science, 9*, 11.

Johnston, M., Dixon, D., Hart, J., Glidewell, L., Schröder, C., & Pollard, B. (2014). Discriminant content validity: A quantitative methodology for assessing content of theory-based measures, with illustrative applications. *British Journal of Health Psychology, 19*, 240-257.

Johnston, M., Carey, R.N., Connell Bohlen, L.E., Johnston, D., Rothman, A., de Bruin, M., Kelly, M.P., Groarke, H. & Michie, S. (2018). Linking behavior change techniques and mechanisms of action: Triangulation of findings from literature synthesis and expert consensus. *Ann. Behav. Med*. preprint DOI: 10.31234/osf.io/ur6kz

Lynn, M. R. (1986). Determination and quantification of content validity. *Nurs Res, 35*, 382-385.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: integrating new and existing techniques. *MIS Quarterly, 35*, 293-334.

Polit , D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health, 29*, 489-497.

Pollard, B., & Johnston, M. (2001). Problems with the Sickness Impact Profile: a theoretically based analysis and a proposal for a new method of implementation and scoring. *Social Science & Medicine, 52*, 921-934.

Pollard, B., Johnston, M., & Dieppe, P. (2006). What do osteoarthritis heath outcome instruments measure? Impairment, activity limitation or participation restriction? *Journal of Rheumatology, 33*, 757-763.

Prinsen, C. A. C., Vohra, S., Rose, M. R., Boers, M., Tugwell, P., Clarke, M., Williamson, P.R. &. Terwee, C. B. (2016). How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" – a practical guideline. *Trials, 17*, 449.

Schmitt, M. A., Schröder, C. D., Stenneberg, M. S., van Meeteren, N. L. U., Helders, P. J. M., Pollard, B., & Dixon, D. (2013). Content validity of the Dutch version of the Neck Bournemouth Questionnaire. *Manual Therapy, 18*(5), 386-389.

Sireci, S. G. (1998). The Construct of Content Validity. *Social Indicators Research, 45*(1), 83-117.

Sireci, S. G., & Sukin, T. (2013). Test validity. In *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology.* (pp. 61-84). Washington, DC, US: American Psychological Association.

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L.M., de Vet, H.C.W. & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research*. 27, 1159-1170

Yalow, E. S., & Popham, W. J. (1983). Content Validity at the Crossroads. *Educational Researcher, 12*, 10-21.

Table 1: Definitions of Validity (adapted from the APA)

Type of Validity	**Definition**
Construct	The degree to which an instrument is capable of measuring a concept, trait, or other theoretical entity. For example, if a researcher develops a new questionnaire to evaluate respondents' levels of aggression, the construct validity of the instrument would be the extent to which it actually assesses aggression as opposed to assertiveness, social dominance, and so forth. There are two main forms of construct validity in the social sciences: **convergent validity and discriminant validity.** **convergent validity:** the extent to which responses on a test or instrument exhibit a strong relationship with responses on conceptually similar tests or instruments. Also called congruent validity. **discriminant validity:** the degree to which a test or measure diverges from (i.e., does not correlate with) another measure whose underlying construct is conceptually unrelated to it. Also called divergent validity.
Criterion	An index of how well a test correlates with an established standard of comparison (i.e., a criterion). Criterion validity is divided into three types: (a) predictive validity, (b) concurrent validity, and (c) retrospective validity. For example, if a measure of health is valid, then it should be possible to use it to predict whether an individual (a) will use health services in future, (b) is currently using health services, and (c) has previously used health services. Also called criterion-referenced validity; criterion-related validity.
Content	The extent to which a test measures a representative sample of the subject matter or behaviour under investigation. For example, if a test is designed to survey health behaviours in older adults, content validity indicates how well it represents the range of health behaviours possible for that population..

Definition 1 Perceived control is the perception of one's capabilities to organise and execute courses of action required to produce given attainments	**Definition 2** Perceived control is the perception of the ease or difficulty of performing the behaviour of interest

How confident are you that you will be able to walk more?

Theoretical Definition	Question measures definition?		How confident are you in your judgement?										
Definition 1	(Yes)	No	0	10	20	30	40	50	60	70	80	(90)	100
Definition 2	Yes	(No)	0	10	20	30	(40)	50	60	70	80	90	100

Figure 1: Illustration of a completed DCV judgement. An item designed to measure self-efficacy is judge against the definition of self-efficacy and perceived behavioural control

Content validity of measures of theoretical constructs in health psychology: discriminant content validity is needed

Introduction

Many of the theoretical constructs and outcomes of interest to health psychology cannot be objectively assessed. For example, phenomena such as beliefs, pain, health, quality of life, stress, intention, illness representations are all of interest but none are available for direct measurement. Rather, the measurement of such theoretical constructs is an inferential process requiring the development of instruments that assess the target construct indirectly, typically using questionnaire based measures. Establishing and reporting the psychometric properties of such measures is challenging but fundamental to their utility in testing theory, designing and evaluating interventions and making clinical and policy decisions.

The psychometric assessment of measures of these health related constructs, including health outcomes and predictors of health outcomes, has advanced in terms of reliability (the degree to which scores on a measure are consistent) and some aspects of construct validity . By contrast, current conventions of reporting typically omit content validity. Examination of the history of the concept of validity sheds some light on why content validity has come to be neglected. Up until the 1980's the APA Standards for Educational and Psychological testing adopted a tripartite approach to (test) validity, namely, criterion, content and construct related validity (Sireci & Sukin, 2013). The current APA definition of each form, adapted to be relevant to health psychology, is given in Table 1. However, this tripartite approach was replaced by a unitary conceptualisation of validity in which construct validity subsumed all other aspects, categories or types of validity. Although the proponents of the unitary conceptualization of validity recognised the importance of representative and relevant content they nonetheless argued that representative and relevant content is not a form of validity. This argument has, over time, likely led to a neglect of content validity. However, the unitary conceptualization, has never been universally accepted. Indeed, some predicted that the refusal to accept content validity as a form of validity would eventually be detrimental to validation practices (Sireci, 1998; Yalow & Popham, 1983); we share that view.

Table 1 about here

In addition, the lack of agreed transparent methods of assessing and reporting content validity also contribute to its neglect. Thus it is possible to report *'how well'* a measure is performing without being able to specify *'what'* it is measuring. Here we argue that the explicit evaluation of content validity would enable researchers and clinicians to select existing measures that truly assess what they aim to measure without ambiguity, overlap or contamination from other related constructs and, where no such measures exist, enable them to develop new content valid measures. By adopting a convention of reporting content validity the pitfalls of poor content validity might be avoided and methods of establishing and assessing content validity might be improved.

Defining content validity and discriminant content validity

Content validity refers to "*the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular purpose*" (Haynes, Richard, & Kubany, 1995). Content validity is considered the most important aspect of a measure of a theoretical construct (Terwee et al., 2018). Content validity is fundamental as it specifies what is being measured. Establishing the content validity of a measure requires that both components, relevance and representativeness, be assessed. First, does the measure accurately reflect the focal construct, i.e. the theoretical construct it aims to measure; do the items that form the measure have relevance to the focal construct? Second, does the measure reflect the whole breadth of the focal construct; is the measure representative of the whole construct? Thus content validity is determined by the relationship between the definition of a construct and the items designed to measure it.

Content validity has the potential to influence the interpretation of all other psychometric properties of a measure, including construct validity and reliability. Most obviously, if a measure is found to be highly reliable but has poor content validity, the interpretation of a score will be entirely erroneous because the measure has no meaning in relation to the target construct.

Construct validity requires that a measure functions as the proposed construct does, but it is possible to achieve construct validity without content validity. For example, in testing the relationship between intention and activity, a measure might predict activity in a manner consistent with theory without containing any intention content, e.g., if it sampled other cognitions related to intention such as attitudes toward activity. And, whilst it is important to establish the internal structure of a measure (Crutzen & Peters,

2017), finding evidence that a measure has the hypothesised factor structure does not in itself demonstrate content validity. Although authors may choose to 'name' the factors to match the intended constructs, the factor may simply contain items which assess the determinants or consequences of the intended construct.

Even when a measure has content validity, it may not have *discriminant content validity*, i.e. content that is distinguishable from the content relevant to other constructs. This can be a significant problem where there are closely related or overlapping constructs. For example, self-efficacy and perceived behavioural control are similar constructs from different theories, but examination of the content validity of measures of these constructs found items which related to neither definition and even one item purporting to measure perceived behavioural control but instead measuring self-efficacy (Johnston et al., 2014). Similarly, measures of pain-catastrophising were considered to have good construct validity due to their performance in predicting activity limitations, but a recent analysis of the content of six standard measures suggest that they lack discriminant content validity. The measures did not adequately reflect the definition of pain catastrophizing and were a better fit to other pain constructs including 'pain-related worrying' or 'pain-related distress' (Crombez, De Paepe, Veirman, Eccleston, & Van Ryckeghem, in submission).

Importance of (discriminant) content validity for theory, intervention design and practice

Content validity is important, for the testing of theory, for the design of behaviour change interventions and for the measurement of health outcomes of importance to patients and clinicians. A lack of content validity weakens theory testing and the interpretation of data.

Valid tests of theory depend on having measures with discriminant content validity, otherwise apparent relationships may simply be due to measurement confound. Contamination of a measure by content relevant to a related construct is particularly problematic when the measures are used to examine relationships between the two constructs. For example, many studies of people with musculoskeletal conditions examine the theoretical relationship between impairments, such as pain, and activity limitations, such as limitations in the ability to walk or get dressed. However, existing outcome measures typically have a mixture of content embracing both impairment and activity limitations (Pollard, Johnston, & Dieppe, 2006). Thus any relationships found may simply be due to the contaminating content, i.e., lack of discriminant content validity in the measures. When pure, uncontaminated measures are used, the relationships are considerably diminished or non-existent (Pollard & Johnston, 2001).

Knowing what one is measuring may be crucial in designing an intervention. If measures lack content validity but are otherwise psychometrically sound, then they may lead to mis-identification of determinants of the target mood, behaviour or symptom. Recent work has identified behaviour change techniques that are appropriate for targeting key theoretical constructs that act as mechanisms of action to determine change in the target behaviour (Carey et al., 2018; Connell et al., 2018; Johnston et al., 2018). Clearly success in making use of such evidence depends on valid labelling of the theoretical construct and this process may be derailed by misleading classification of measures. For example, some self-efficacy items have been found to tap 'motivation' as well as self-efficacy (Burrell, Allan, Williams, & Johnston, 2018) and if used as a basis for designing an intervention might result in the selection of less appropriate techniques than might be achieved if determinants of the behaviour were identified using measures with content validity.

In practical applications, test scores are used to make decisions, including planning and policy decisions that affect availability of resources, as well as decisions about clinical interventions. If hospital managers attempting to reduce work stress in nurses used current measures to assess factors influencing work stress, they might be misled about the key determinants, such as workers' control, since the main measure assesses both control over decisions and use of skill and individual items have poor content validity (Bell, Johnston, Allan, Pollard, & Johnston, 2017). A clinician might reach a different treatment decision depending on the content validity of the measure used for patient assessment. Measures used to assess limitations in patients with arthritis range from those which contain items mainly assessing impairment to those which are mainly participation restrictions (Pollard et al., 2006); it is therefore unsurprising if they result in differing assessments of degrees of severity and if they differ in sensitivity to treatment effects of pharmacological and exercise interventions (Ayis et al., 2010).

Developments of digital and mobile technologies make the content validity problem more, rather than less, urgent. With the exception of direct observational and sensing measures, the commonly used EMA (ecological momentary assessment) methods rely on very frequent self-reporting of behaviours or of intrapsychic phenomena such as beliefs, emotions, symptoms and their precursors or consequences. In order to reduce participant burden, each construct is typically measured by a single or a few items and it is therefore particularly important that each of these items has content validity for its target construct.

The use of an agreed, feasible, evidence-based methodology for establishing essential aspects of content validity of construct measures, would considerably advance our ability to develop and/or select valid measures suitable for theory testing and intervention development, and the assessment of outcomes important to patients and clinicians.

Methods of assessing CV and DCV

Two methods are currently available for assessing content validity, namely, the Content Validity Index (CVI) (Lynn, 1986) and Discriminant Content Validity (DCV) (Johnston et al., 2014). While CVI has been used to assess content validity of health outcome measures, DCV has been used to assess the content validity of theoretical process variables (Bell et al., 2017; Burrell et al., 2018; Gardner, Abraham, Lally, & de Bruijn, 2012; Johnston et al., 2014) and theoretical domains (Cane, O'Connor, & Michie, 2012; Huijg, Gebhardt, Crone, Dusseldorp, & Presseau, 2014), as well health outcome measures (Dixon, Johnston, Mcqueen, & Court-Brown, 2008; Dixon, Pollard, & Johnston, 2007; Pollard et al., 2006; Schmitt et al., 2013).

Both are judgement tasks, whereby judges rate the extent to which measurement items are related to the target construct. Therefore, both CVI and DCV can be used in the development of a valid measure, prior to evaluation of reliability and construct validity by testing with participants. However, there are differences between the two methods. CVI examines the relevance of items in relation to a single target construct and is scored such that each item is judged either relevant or not. The proportion of judges in agreement about relevance is then used to determine the CVI for each item (I-CVI) and the CVI for the whole scale/measure (S-CVI) (Polit & Beck, 2006). However, it does not quantify the extent to which the measure is distinct and uncontaminated by other constructs.

Figure 1 about here

We developed the DCV method to provide that additional important information. Like the CVI, DCV establishes the content validity of items in relation to the target construct, but also establishes content validity in relation to other constructs, either from the same theory or related constructs from other theories. In addition, a DCV study also identifies items that *do not* measure the construct (see the illustration in Figure 1, where the item is judged, with high confidence, to be measure self-efficacy, but with more modest confidence to *not* measure perceived behavioural control). Thus, DCV establishes the discriminant content validity of items necessary for theory testing and for precise assessment of intervention effects. DCV data

also give a quantitative estimate of the content validity of items (Crombez et al., in submission; Johnston et al., 2014). Thus, DCV can be used to identify items that are pure measures of a single construct uncontaminated by other constructs and can examine whether an item measures one construct more strongly than others. It is then possible to choose items as required, for example, theory testing requires pure measures, whilst items measuring multiple constructs may be useful as single item measures, especially in clinical settings (Johnston et al., 2014).

Further challenges

Both CVI and DCV methods focus on 'relevance' and only one DCV study has attempted to develop a method for assessing representativeness (Bell et al., 2017), the other key component of content validity. Standard measures of work stress lacked items to assess important parts of work stress definitions, and also contained items that were not relevant to the definitions. Similar attempts to assess representativeness have been explored in organizational psychology (MacKenzie, Podsakoff, & Podsakoff, 2011).

Content validity studies have highlighted the importance of the definitions of the target constructs and others have noted this problem in establishing scale validity: "Given the importance of clearly defining the conceptual domain of the construct, it is surprising that so many researchers either neglect this step in the process or fail to properly implement it" (MacKenzie et al., 2011). If definitions are imprecise or simply lacking then it is virtually impossible to establish content validity. For example, there are no agreed definitions of work stress constructs and the content validity of items in a self-report measure of work stress differs when different definitions of the construct are used (Bell et al., 2017).

To date, content validity studies have tended to focus on theoretical constructs and health outcome measures with rather less emphasis on the content validity of other types of constructs, such as process variables and other outcomes of interest to health psychology, for example, self-reported measures of adherence, dietary or other health behaviours. Most of the work has been done on the content of the question and less on the content of the response format: for example, evaluating frequency versus intensity versus agreement formats. Future work might usefully begin to apply content validation methods more widely.

In addition, both CVI and DCV are judgment tasks, yet there is no agreement as to how judges should be selected. Typically, expert judges are used but should judges be experts familiar with the theoretical constructs or should they be drawn from the population of participants who will be the

respondents when the measure is used? The performance of expert judges on a DCV task were found to have clearer discriminations of items assessing illness perceptions than lay respondents similar to the intended respondents using the measures (Glidewell, 2008). Thus items that might have CV for the expert judges might lack validity for lay judges. This raises the issue of synthesis across content validity studies. There is surely merit in replication studies that examine the content validity of the same items or measures using different samples of the same type of judge or across different types of judges. Suitable methods of data aggregation could then be used to improve the reliability of content validity measures.

Finally, neither method overcomes the need for the robust methods to initially establish the pool of items relevant to, and representative of, the target construct. Qualitative methods including focus groups, interviews, and cognitive interviews with the target population have been used, together with expert opinion to improve the range and intelligibility of items (Prinsen et al., 2016).

Conclusions

Good measurement of key theoretical constructs is fundamental to much research and application in health psychology and we typically present careful assessments of aspects of reliability and validity. However we neglect the need to demonstrate the content validity of our measures. Without satisfactory content validity and discriminant content validity results may be meaningless and worse, may lead to erroneous conclusions in testing theory, choosing interventions and making clinical and policy decisions.

In part this deficit may be due to lack of clearly established methods, but CVI has been available for many years. It would therefore appear that we have simply established a convention of omission of content validity as part of the reporting of psychometric properties. There has been some recognition of the confusion of overlapping, ambiguous and confounding of labels for theoretical constructs but we have avoided contemplation of the hazards of failure to establish the discriminant content validity of our measures. We suggest that in future authors, reviewers and editors might seek better evidence of content validity (and especially discriminant content validity) of measures used in empirical studies and that research continues to address the challenges of improving content validity assessment.

References

Ayis, S., Arden, N., Doherty, M., Pollard, B., Johnston, M., & Dieppe, P. (2010). Applying the Impairment, Activity Limitation and Participation Restriction Constructs of ICF (WHO) Model to OA and low back pain trials: re-analysis *Journal of Rheumatology, 37*, 1923-1931.

Bell, C., Johnston, D., Allan, J., Pollard, B., & Johnston, M. (2017). What do Demand-Control and Effort-Reward work stress questionnaires really measure? A discriminant content validity study of relevance and representativeness of measures. *British Journal of Health Psychology, 22*, 295-329.

Burrell, A. M. G., Allan, J. L., Williams, D. M., & Johnston, M. (2018). What do self-efficacy items measure? Examining the discriminant content validity of self-efficacy items. *British Journal of Health Psychology*, 23: 597-611.

Cane, J., O'Connor, D., & Michie, S. (2012). Validation of the theoretical domains framework for use in behaviour change and implementation research. *Implementation Science, 7*, 37.

Carey, R. N., Connell, L. E., Johnston, M., Rothman, A. J., de Bruin, M., Kelly, M. P., & Michie, S. (2018). Behavior Change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. *Ann Behav Med*. kay078, https://doi.org/10.1093/abm/kay078.

Connell, L. E., Carey, R. N., de Bruin, M., Rothman, A. J., Johnston, M., Kelly, M. P., & Michie, S. (2018). Links between behavior change techniques and mechanisms of action: an expert consensus study. *Ann Behav Med*. doi:10.1093/abm/kay082

Crombez, G., De Paepe, A. L., Veirman, E., Eccleston, C., & Van Ryckeghem, D. (in submission). Let's talk about pain catastrophizing measures: an item content analysis. *Pain*.

Crutzen, R., & Peters, G. Y. (2017). Scale quality: alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychol Rev, 11*, 242-247.

Dixon, D., Johnston, M., Mcqueen, M., & Court-Brown, C. (2008). The Disabilities of the arm, shoulder and hand questionnaire (DASH) can measure the impairment, activity limitations and participation restriction constructs from the international classification of functioning, Disability and Health (ICF). *BMC Musculoskeletal Disorders, 9*.

Dixon, D., Pollard, B., & Johnston, M. (2007). What does the chronic pain grade questionnaire measure? *Pain, 130*(3), 249-253.

Gardner, B., Abraham, C., Lally, P., & de Bruijn, G. J. (2012). Towards parsimony in habit measurement: testing the convergent and predictive validity of an automaticity subscale of the Self-Report Habit Index. *Int J Behav Nutr Phys Act, 9*, 102.

Glidewell, E. (2008). *How do people with a diagnosis, caregivers and healthcare professionals represent dementia: An exploratory assessment using the Common Sense Self Regulation Model.* University of Aberdeen, PhD Thesis.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*, 238-247.

Huijg, J. M., Gebhardt, W. A., Crone, M. R., Dusseldorp, E., & Presseau, J. (2014). Discriminant content validity of a theoretical domains framework questionnaire for use in implementation research. *Implementation Science, 9*, 11.

Johnston, M., Dixon, D., Hart, J., Glidewell, L., Schröder, C., & Pollard, B. (2014). Discriminant content validity: A quantitative methodology for assessing content of theory-based measures, with illustrative applications. *British Journal of Health Psychology, 19*, 240-257.

Johnston, M., Carey, R.N., Connell Bohlen, L.E., Johnston, D., Rothman, A., de Bruin, M., Kelly, M.P., Groarke, H. & Michie, S. (2018). Linking behavior change techniques and mechanisms of action: Triangulation of findings from literature synthesis and expert consensus. *Ann. Behav. Med*. preprint DOI: 10.31234/osf.io/ur6kz

Lynn, M. R. (1986). Determination and quantification of content validity. *Nurs Res, 35*, 382-385.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: integrating new and existing techniques. *MIS Quarterly, 35*, 293-334.

Polit , D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health, 29*, 489-497.

Pollard, B., & Johnston, M. (2001). Problems with the Sickness Impact Profile: a theoretically based analysis and a proposal for a new method of implementation and scoring. *Social Science & Medicine, 52*, 921-934.

Pollard, B., Johnston, M., & Dieppe, P. (2006). What do osteoarthritis heath outcome instruments measure? Impairment, activity limitation or participation restriction? *Journal of Rheumatology, 33*, 757-763.

Prinsen, C. A. C., Vohra, S., Rose, M. R., Boers, M., Tugwell, P., Clarke, M., Williamson, P.R. &. Terwee, C. B. (2016). How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" – a practical guideline. *Trials, 17*, 449.

Schmitt, M. A., Schröder, C. D., Stenneberg, M. S., van Meeteren, N. L. U., Helders, P. J. M., Pollard, B., & Dixon, D. (2013). Content validity of the Dutch version of the Neck Bournemouth Questionnaire. *Manual Therapy, 18*(5), 386-389.

Sireci, S. G. (1998). The Construct of Content Validity. *Social Indicators Research, 45*(1), 83-117.

Sireci, S. G., & Sukin, T. (2013). Test validity. In *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology.* (pp. 61-84). Washington, DC, US: American Psychological Association.

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L.M., de Vet, H.C.W. & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research*. 27, 1159-1170

Yalow, E. S., & Popham, W. J. (1983). Content Validity at the Crossroads. *Educational Researcher, 12*, 10-21.